IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 10, May 2023

Review on Different Types of Algorithm for Data Mining

Mayur Khandve[1], Ashish Chikne[2], Ganesh Mendhkar[3], Ajnkya Salunkhe[4], Dnyaneshwar Aher[5], Prof Ekta Patel[6]

Students, Department of Computer Engineering[1,2,3 4,5]

Professor, Department of Computer Engineering[5]

Shree Ramchandra College of Engineering Pune, Maharashtra, India

Abstract: *This review report explores the application of data mining algorithms in student performance analysis. It provides an overview of the significance of student performance analysis in education and highlights the role of data mining techniques in extracting valuable insights from educational data. The report reviews various data mining algorithms used in student performance analysis, examines their strengths and limitations, discusses relevant studies in the field, and provides insights for future research directions.*

Keywords: Linear regression, Predictive modelling, Educational data mining, Academic outcomes, Intervention strategies, Data pre-processing, Feature selection;

I. INTRODUCTION

In the field of education, the analysis of student performance plays a vital role in understanding student outcomes, identifying at-risk students, and designing effective intervention strategies. With the proliferation of educational data and advancements in data mining techniques, researchers and educators have increasingly turned to data mining algorithms to gain valuable insights from educational datasets. This review report aims to provide an in-depth exploration of the application of data mining algorithms in student performance analysis.

The significance of student performance analysis in education cannot be overstated. By analysing various factors such as previous grades, demographics, socio-economic background, and learning behaviours, educators can gain insights into the factors that influence student outcomes. This knowledge enables them to personalize educational strategies, identify students who may be at risk of academic failure, and implement timely interventions to improve student success rates.

1.1 Abbreviation and Acronyms

- SPA - Student Performance Analysis
- DM - Data Mining
- LR - Logistic Regression
- DT - Decision Trees
- LOR - Linear Regression

1.2 Data Mining:

Data mining is a multidisciplinary field that involves the discovery of patterns, relationships, and insights from large datasets. It encompasses a wide range of techniques, algorithms, and methodologies for extracting valuable knowledge from data. The ultimate goal of data mining is to transform raw data into meaningful and actionable information that can be used for decision-making, prediction, and knowledge discovery. [1]

Process of Data Mining:

The process of data mining typically involves several steps:

DOI: 10.48175/IJARSCT-10452

- Data Collection: Gathering and assembling relevant data from various sources, such as databases, data warehouses, or external repositories.
- Data Pre-processing: Cleaning and transforming the collected data to ensure its quality, consistency, and suitability for analysis. This step involves handling missing values, removing noise and outliers, and standardizing or normalizing data.
- Exploratory Data Analysis (EDA): Conducting an initial exploration and visualization of the data to gain insights, identify patterns, and detect relationships or anomalies. EDA helps in understanding the characteristics of the dataset and guides subsequent analysis.
- Feature Selection/Engineering: Identifying the most relevant features or variables that have the most significant impact on the target variable. Feature selection aims to reduce dimensionality and improve the efficiency and interpretability of the data mining process.

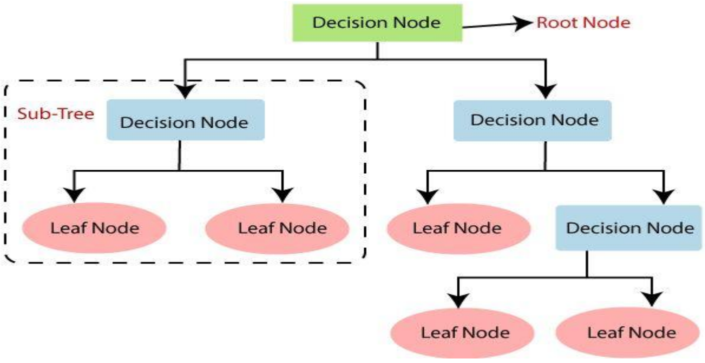
Advantages of Data Mining

- Personalized Education: Data mining allows for the development of personalized education strategies. By analysing individual student data, educators can tailor instruction, interventions, and support to meet the specific needs of students, enhancing their learning experiences.
- Early Identification of At-Risk Students: Data mining algorithms can help identify students who are at risk of academic failure or dropping out. Early identification allows for timely intervention and support, potentially improving student outcomes and reducing dropout rates.
- Disadvantages of Data Mining:
- Data Quality and Availability: Data mining heavily relies on the quality and availability of data. Incomplete, inconsistent, or inaccurate data can affect the performance and reliability of data mining algorithms, leading to biased or misleading results.
- Interpretability and Explain ability: Some advanced data mining algorithms, such as neural networks or ensemble methods, may lack interpretability. Understanding and explaining the underlying reasons behind the predictions or patterns identified by these algorithms can be challenging.

1.3 Algorithms of Data Mining

A. Decision tree

Explanation: Decision trees are tree-like models that use a hierarchical structure of nodes and branches to make predictions or decisions. Each internal node represents a test on a specific attribute, and each leaf node represents a class or outcome. [2]



Advantages:

- Easy to understand and interpret.
- Can handle both categorical and numerical data.
- Can capture nonlinear relationships.

DOI: 10.48175/IJARSCT-10452

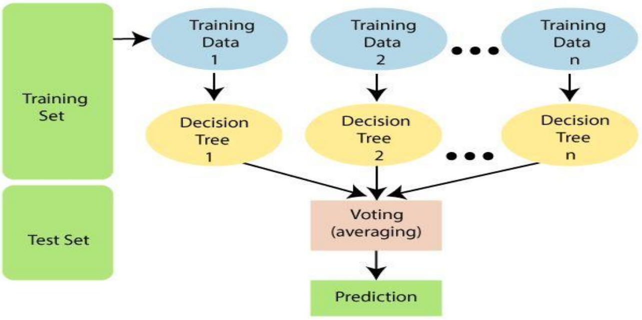
- Can handle missing values and outliers.

Disadvantages:

- Prone to over-fitting, especially with complex and deep trees.
- Can be sensitive to small changes in the data, leading to different tree structures.
- Limited in handling irrelevant attributes or redundant data.

B. Random Forest

Random Forests is an ensemble learning technique that combines multiple decision trees to make predictions. It creates an ensemble of decision trees by training each tree on a randomly sampled subset of the training data and using majority voting or averaging for predictions. [3]



Advantages:

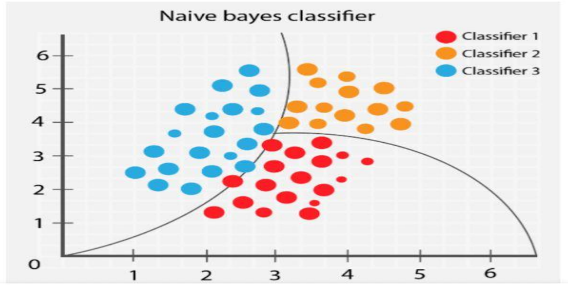
- Handles high-dimensional data effectively.
- Provides feature importance rankings.
- Robust to outliers and missing data.

Disadvantages:

- More complex than individual decision trees, making interpretation challenging.
- Can be computationally expensive, especially for large datasets and many trees.
- Requires careful tuning of hyper-parameters to optimize performance.

C. Naive Bayes Algorithm

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between predictors. It calculates the probability of a particular class given the values of the predictors and selects the class with the highest probability. [4]



Advantages:

- Fast and computationally efficient.
- Performs well with high-dimensional data.
- Handles missing data gracefully.
- Works well with categorical predictors.

DOI: 10.48175/IJARSCT-10452

Disadvantages:

- Assumes independence between predictors, which may not hold in real-world scenarios.
- Sensitive to the presence of irrelevant or correlated predictors.
- Limited in capturing complex relationships.

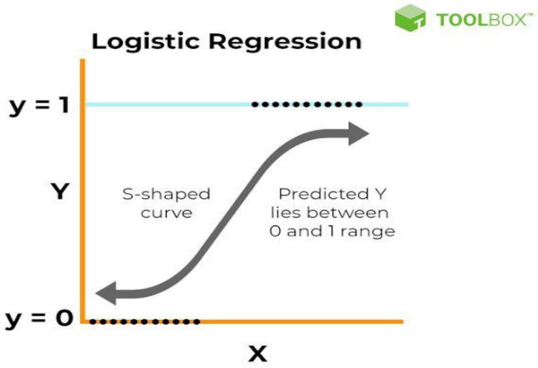
D. Logistic Regression:

Explanation: Logistic regression is a classification algorithm used when the dependent variable is binary (e.g., pass/fail). It estimates the probability of a student belonging to a particular class based on the independent variables. [5]

Advantages:

- Probabilistic interpretation of the results.
- Can handle both numerical and categorical independent variables.
- Relatively simple and efficient to implement.

1.



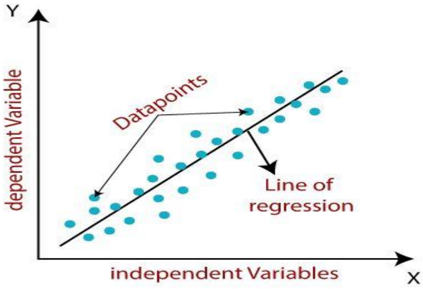
Disadvantages:

- Assumes a linear relationship between independent variables and the log-odds of the outcome, which may not hold true.
- Cannot handle non-linear relationships without feature engineering.
- May suffer from over-fitting or under-fitting if the model is too complex or too simple, respectively.

E. Linear regression

Linear regression is a statistical modelling technique used to establish a linear relationship between a dependent variable (student performance) and one or more independent variables (predictor variables) such as previous grades, attendance, or study time. It predicts the numerical value of the dependent variable based on the values of the independent variables.

Linear regression is a well-established and widely used technique in the field of predictive analytics. It has a strong theoretical foundation and is extensively studied, making it easier to interpret and understand the results. This interpretability is crucial in educational settings, where educators need to comprehend and explain the factors influencing student performance. [6]



Advantages

- Simple and easy to interpret.
- Provides information about the magnitude and direction of relationships between variables.
- Can handle both continuous and categorical independent variables.

Disadvantages

- Assumes a linear relationship between variables, which may not always hold true.
- Sensitive to outliers that can significantly affect the model's performance.
- Limited in capturing complex nonlinear relationships.

II. CONCLUSION

In this paper, we studied different types of algorithm used for data mining such as linear regression, logistic regression, decision trees, random forests, and naive Bayes. From this studied we conclude that decision tree algorithm is easy to understand and interpreted but very sensitive to small changes in the data, random forest algorithm handles high dimensions data effectively but more complex and expensive foe large dataset, naïve Bayes algorithm fast and efficient but limited in capturing complex relationship, logistic regression algorithm handles both numerical and categorical variables effectively but may suffer from over-fitting and under-fitting if the model is to complex or to simple, linear regression algorithm has many advantages like simple and easy to interpret, handle all types of independent variables and provide information about variables due to this algorithms linear regression algorithm can be best suited for analysis of student performance application

REFERENCES

[1]. Qingqing Chang,Research and Application of the Data Mining Technology in Economic Intelligence System,Volume 2022 | Article ID 6439315 ,Cognitive Computing Paradigms for Medical Big Data Processing and its Trends| https://doi.org/10.1155/2022/6439315.

[2]. Chenmeng Zhang,Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation,Journal of Physics: Conference Series ,1732 (2021) 012086 IOP Publishing doi:10.1088/1742-6596/1732/1/012086.

[3]. Falguni Ghatkar, Sakshi kharche, Priyanka Doifode, Jagruti Khairnar, Prof. Neelam Kumar," To Study Different Types of Supervised Learning Algorithm" May 2023, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue 8, May 2023,PP-25-32

[4]. Hong Chen,"Improved naive Bayes classification algorithm for traffic risk management",EURASIP Journal on Advances in Signal Processing volume 2021, Article number: 30 (2021), Published: 22 June 2021

[5]. CHAO-YING JOANNE PENG,"An Introduction to Logistic Regression Analysis and Reporting",The Journal of Educational Research,2002,PP1-14

[6]. Kirti Gadekar ,"GENERAL LINEAR REGRESSION",2018 IJRAR January 2019, Volume 06, Issue 1 www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138), PP42-46

[7]. Durgesh Ugale, JeetPawar, SachinYadav,Dr.Chandrashekhar Raut, "Student Performance Prediction Using Data Mining Techniques", international research journal of engineering and technology(irjet)volume:07issue:05|may2020.

[8]. Somya Mishra,Mrunal Lokare, Aniket Patil, Prof.Chandrashekhar Badgujar, "Student Performance Analysis System", internationalresearchjournalofengineeringandtechnology(irjet)volume:08 issue:04|apr2021.

[9]. SyedFurkhanMehdi,SyedRayyan,MohdYaseen,RafathSamrin"StudentPerformanceAnalysisBasedMachineLearning"journalofinformationandcomputationalscienceissn:1548-7741

DOI: 10.48175/IJARSCT-10452